New York Scientific Data Summit (NYSDS)

PROGRAM

August 7-9, 2017

New York University, New York, NY 10012
https://www.bnl.gov/nysds17

BROOKHAVEN NATIONAL LABORATORY
Stony Brook University
NYU
iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE
IEEE computer society

Welcome to NYSDS 2017! We are delighted that you are able to attend and participate in this year's summit. The objective of this conference series is to accelerate data-driven discovery and to build communities and collaborations by bringing together leading researchers, developers and end-users from academia, industry, utilities and state and federal governments.

This year's theme is again "Data-Driven Discovery in Science and Industry" and we are privileged to have such excellent speakers from a broad range of organizations in industry, academia and national laboratories. The summit will focus on 4 different themes this year, which are at the forefront of many data discovery challenges:

- **Streaming Data Analysis** – The ability to extract knowledge and support decision making in environments, where data arrives at high speeds and volumes, and the response time is limited.
- **Autonomous Experimental Design and Optimization** – Supporting complex and/or high throughput experiments by utilizing AI and Computational Modeling to plan and adapt the experiments in order to optimize the scientific outcome.
- **Performance for Big Data** – Addressing the performance challenges posed by big data applications from the hardware to the application level.
- **Interactive Exploration of Extreme Scale Data** – Ability to support scientific discovery – interactive, collaborative – in data sets exceeding multiple PBs in size, such as are found, in particular, in nuclear physics, high energy physics, computational biology and climate science.

We are looking forward to many interesting discussions around these four themes.

Led by the Computational Science Initiative (CSI) at Brookhaven National Laboratory (BNL), the summit is jointly organized by BNL, Stony Brook University (SBU), New York University (NYU) and the IEEE Computer Society (Long Island and New York Chapter). A special thank you goes to our NYU hosts whose wonderful facilities we are using, to our program committee for assembling the program, and to our local organizing committee responsible for the smooth running of the event.

Thanks are also due to our sponsors – their support is crucial to the success of this event and we also acknowledge them as essential partners or technology providers in our research endeavors.

Finally, we would like to wish you all well and hope that you have a great time both in the meeting and out exploring New York City.

Kerstin Kleese van Dam

Director of Brookhaven National Laboratory's Computational Science Initiative

COMMITTEES

Program Committee

Kerstin Kleese van Dam (BNL)

Frank Alexander (BNL)

Barbara Chapman (SBU/BNL)

Nicholas D'Imperio (BNL)

Michael McGuigan (BNL)

Lauri Peragine (BNL)

Kyle Cranmer (NYU)

Robert Harrison (SBU/BNL)

Metodi Filipov (IEEE)

Marjaneh Issapour (IEEE)

Local Organizing Committee

Lauri Peragine (BNL)
Gina Liles (BNL)
Eileen Pinkston (BNL)
Maureen Anderson (BNL)
Michael McGuigan (BNL)

Co-hosts

Brookhaven National Laboratory
Stony Brook University
NYU Center for Data Science
Institute for Advanced Computational Science
The Moore-Sloan Data Science Environment
IEEE Computer Society-Long Island Chapter

The Organizing Committee would like to thank our sponsors for making our third annual New York Scientific Data Summit (NYSDS) a success.

AGENDA: Monday, August 7th

SESSION 1: STREAMING DATA ANALYSIS

Location: Eisner & Lubin Auditorium (4th floor)

Kerstin Kleese van Dam, Chair
Director of BNL'S Computational Science Initiative (CSI)

TIME	SPEAKER	TITLE
8:45-9:00 am	WELCOME REMARKS: Kerstin Kleese van Dam, Director of BNL'S Computational Science Initiative (CSI)	
9:00-9:45 am	**KEYNOTE: Kevin Yager, Brookhaven National Laboratory (BNL)**	**"Exploiting Deep Learning for Automated Synchrotron Experiments"**
9:45-10:20 am	John Wu, Lawrence Berkeley National Laboratory (LBNL)	"Statistical Data Reduction for Streaming Data"
10:20-10:50 am	*Break*	
10:50-11:25 am	Jun Wang, Stony Brook University (State University of New York)	"Incremental Clustering of Big Data with GPU Acceleration and Visualization"
11:25-12:00 pm	Michael DePhillips, BNL	"A Case for High-Bandwidth Monitoring"
12:00-12:35 pm	Yuzhong Yan, Prairie View A & M University	"Implementing a Distributed Volumetric Data Analytics Toolkit on Apache Spark"
12:35-2:00 pm	*Lunch on your own*	
2:00-2:35 pm	Chiwoo Park, Florida State University	"In situ Analytics of High Frame-rate Image Streaming"
2:35-3:10 pm	Nikolay Malitsky, BNL and Aashish Chaudhary, Kitware	"Building Near-Real-Time Processing Pipelines with the Spark-MPI Platform"

SESSION 2: AUTONOMOUS EXPERIMENTAL DESIGN AND OPTIMIZATION

Frank Alexander, Chair
Deputy Director of BNL'S Computational Science Initiative (CSI)

TIME	SPEAKER	TITLE
3:10-3:55 pm	**KEYNOTE: Shantenu Jha, Rutgers University**	**"Building Blocks for Adaptive Workflows"**
3:55-4:25 pm	*Break*	
4:25-5:00 pm	Kristofer Reyes, New York State University at Buffalo	"Closed-loop Autonomous Research Systems (ARES)"
5:00-5:35 pm	Michael McKerns, SBU	"Is Automated Materials Design and Discovery Possible?"
	Dinner on your own	

SESSION 3: PERFORMANCE FOR BIG DATA

Location: Eisner & Lubin Auditorium (4th floor)

Barbara Chapman, Chair

Director of BNL'S Computer Science and Mathematics Department,
Professor of Applied Mathematics & Statistics and Computer Science at Stony Brook University (SBU)

TIME	SPEAKER	TITLE
9:00-9:45 am	**KEYNOTE: Peter Beckman, Argonne National Laboratory (ANL)**	**"The Convergence of Extreme Computing and Big Data: From Edge Computing to Exascale"**
9:45-10:20 am	Hamid Reza Assadi, SBU	"Comparative Study of Deep Learning Framework in HPC Environments"
10:20-10:55 am	*Break*	
10:55-11:30 am	Eric Stephan, Pacific Northwest National Laboratory (PNNL)	"A Scientific Data Provenance Harvester for Distributed Applications"
11:30-12:05 pm	James Jeffers, Intel	"Improving Large Scale Visual Data Analysis using Intel Supported Software Defined Visualization Solutions"
12:10-1:40 pm	*Lunch on your own*	
1:45-2:20 pm	Sameera Abeykoon, BNL	"Parallelizing X-ray Photon Correlation Spectroscopy Software Tools Using Python Multiprocessing"
2:20-2:55 pm	Geoffrey Fox, Indiana University	"A Tale of Two Convergences: Applications and Computing Platforms"
2:55-3:25 pm	*Break*	
3:25-4:00 pm	Zichao (Wendy) Di, ANL	"Multigrid Approach for Tomographic Reconstruction"
4:00-4:35 pm	Line Pouchard, BNL	"Capturing Provenance as a Diagnostic Tool for Workflow Performance Evaluation and Optimization"
4:35-5:10 pm	Ryan Quick, Providentia Worldwide	"Event Stream Optimization for Big Data Stream Analytics"
5:30-6:45 pm	*Poster Session*	Location: Room 914 (9th Floor)
7:00-9:00 pm	*Dinner*	Location: Rosenthal Pavilion (10th Floor)
	KEYNOTE, DINNER SPEAKER: Peter Coveney, University College London	**"Big Theory for Big Data"**

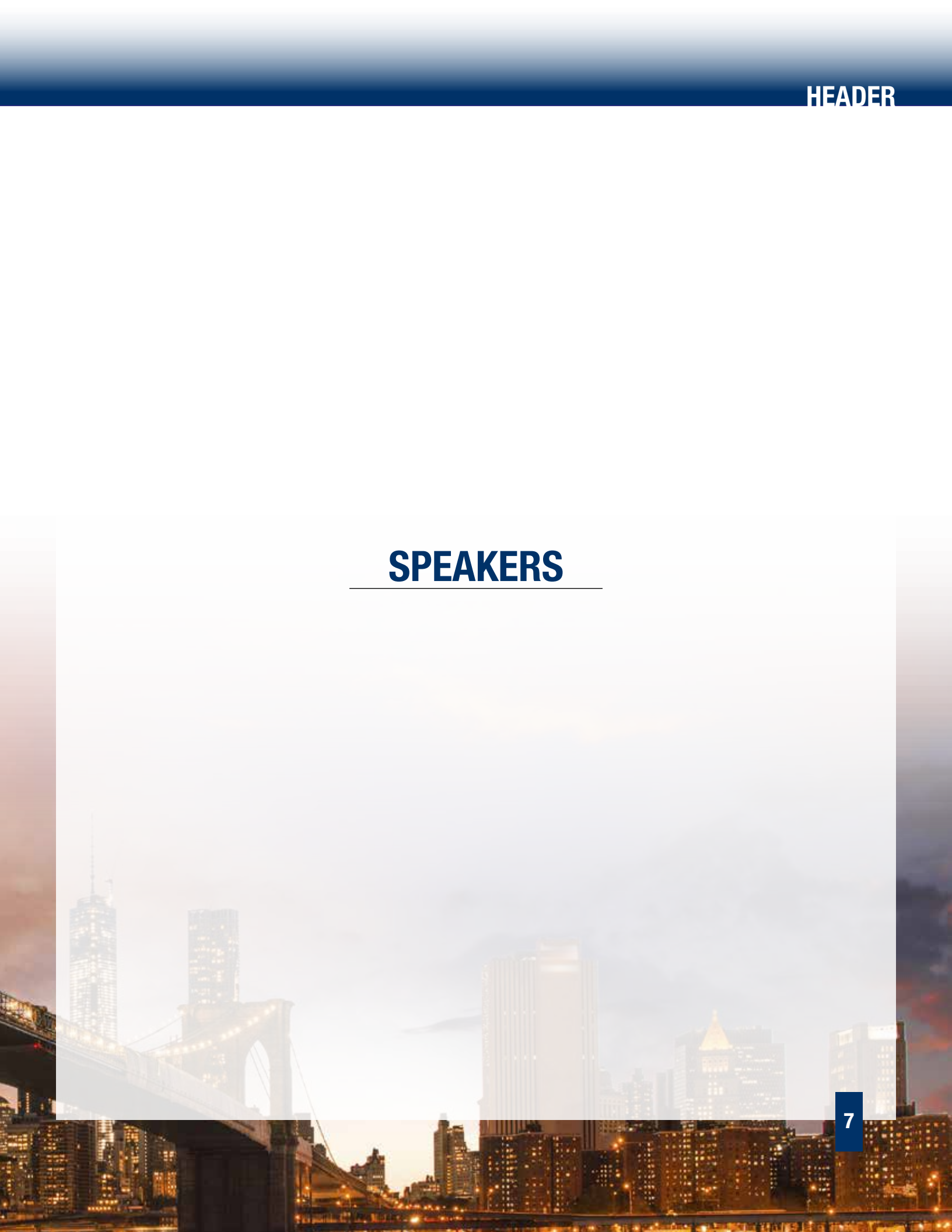
SESSION 4: EXTREME SCALE DATA

Location: Eisner & Lubin Auditorium (4th floor)

Nicholas D'Imperio, Chair
Director of BNL'S Computational Science Laboratory

TIME	SPEAKER	TITLE
9:00-9:45 am	**KEYNOTE: James Ahrens, Los Alamos National Laboratory (LANL)**	**"Supercharging the Scientific Process via Data Science at Scale"**
9:45-10:20 am	Jialin Liu, National Energy Research Scientific Computing /LBNL	"Searching for Millions of Objects in the BOSS Spectroscopic Survey Data with H5Boss"
10:20-10:50 am	*Break*	
10:50-11:25 am	Dantong Yu, New Jersey Institute of Technology	"Robust and Scalable Deep Learning for X-ray Synchrotron Image Analysis"
11:35 am	CLOSING REMARKS: Robert J. Harrison, Director of Institute for Advanced Computational Science (IACS), and Chief Scientist at BNL, CSI	

SPEAKERS



"Exploiting Deep Learning for Automated Synchrotron Experiments"

Kevin Yager - KEYNOTE

Group Leader, Center for Functional Nanomaterials, Brookhaven National Laboratory

Abstract

Modern scientific instruments generate data at unprecedented rates. Manual sorting, classification, and analysis of these data streams is impractical. Moreover, switching to automated streaming data analysis opens the door to innovative autonomous experiments, wherein the scientific instrument intelligently explores parameter spaces without human intervention. This talk will present recent progress in applying machine-learning methods to the classification of x-ray scattering datasets generated at synchrotron beamlines. We demonstrate how deep learning networks can be augmented by taking advantage of the known physics of the scientific problem, yielding vastly improved performance.

Bio

Kevin Yager is the group leader for the Electronic Nanomaterials group in the Center for Functional Nanomaterials at Brookhaven National Laboratory (CFN, BNL). He obtained his Ph.D. in 2006 from McGill University, Department of Chemistry, on photo-responsive polymers, and worked for 3 years as a guest researcher in the Polymers Division at the National Institute of Standards and Technology (NIST) developing neutron scattering methods. His current work at BNL focuses on self-assembled nanostructures, and structural characterization using x-ray scattering. Over the last few years, he has been part of a team designing and building new high-performance x-ray scattering instruments at BNL's National Synchrotron Light Source II (NSLS-II), including developing the hardware and software infrastructure necessary for autonomous experimentation.

"Statistical Data Reduction for Streaming Data"

John Wu
Senior Computer Scientist, Lawrence Berkeley National Laboratory

D. Lee
Lawrence Berkeley National Laboratory
Texas A&M University, Commerce

A. Sim
Lawrence Berkeley National Laboratory

J. Choi
Lawrence Berkeley National Laboratory
Ulsan National Institute of Science and Technology

Abstract

Bulk of the streaming data from scientific simulations and experiments consists of numerical values, and these values often change in unpredictable ways over a short time horizon. Such data values are known to be hard to compress, however, much of the random fluctuation is not essential to the scientific application and could therefore be removed without adverse impact. We have developed a compression technique based on statistical similarity that could reduce the storage requirement by over 100-fold while preserve prominent features in the data stream. We achieve these impressive compression ratios because most data blocks have similar probability distribution and could be reproduced from a small block. The core concept behind this work is the exchangeability in statistics. To create a practical compression algorithm, we choose to work with fixed size blocks and use Kolmogorov-Smirnov test to measure similarity. The resulting technique could be regarded as a dictionary-based compression scheme.

"Incremental Clustering of Big Data with GPU Acceleration and Visualization"

Jun Wang
Research Assistant, Stony Brook University

E. Papenhausen, B. Wang, S. Ha, K. Mueller
Stony Brook University

A. Zelenyuk
Pacific Northwest National Laboratory

Abstract

The availability of big data has been revolutionizing scientific research. Nature and Science have even published special issues dedicated to discuss the opportunities and challenges brought by big data. Nevertheless, the unprecedented growth of data also makes the acquisition of data labels impossible. It makes clustering an unavoidable first step in big data processing. Clustering not only serves as a viable means to reveal the significant nature of the data under investigation, but it is also a necessary step for many possible further analytics, e.g. hierarchical indexing and stratified sampling. However, common clustering techniques, even when GPU accelerated, mostly fail in the big data context due to issues with scalability and the difficulty with handling streaming data. All this makes scalable incremental clustering of large-scale data a so far unsolved challenge. We have tackled this challenge via a novel algorithm which parallelizes incremental K-means. The CPU version of the algorithm runs in an incremental way such that data points are read sequentially and each new point is compared to all existing clusters to see if it belongs to a certain cluster or should be recognized as a new cluster. The computation of the distance matrix can be easily sped up with the GPU.

"A Case for High-Bandwidth Monitoring"

Michael DePhillips
Senior Technology Architect, Brookhaven National Laboratory

D. Katramatos
Brookhaven National Laboratory

Abstract

In this talk we present an effort to architect, research, develop, and test a next-generation, high-bandwidth network monitoring framework designed to handle the rigors of large scientific data feeds. This framework will be capable of transparently capturing and analyzing network traffic in real time so as to enable early and rapid response to potential threats. We seek to adapt and integrate existing and ongoing work on streaming data analysis on the wire and packet capture with real-time analytics using accelerators to create a next-generation, high-bandwidth net- work monitoring framework. Flow interrogation in real time will transparently divert selected network flows to an attached computing infra- structure and subject them to processing and analysis. With acceptable QOS this system will detect suspicious activities, with "innocent" flows allowed to proceed to their original desti- nation and suspicious flows either dropped or further processed and monitored, with appropriate storage and analysis.

"Implementing a Distributed Volumetric Data Analytics Toolkit on Apache Spark"

Yuzhong Yan
Research Associate, Prairie View A&M University

C. Chen, L. Huang, L. Qian
Prairie View A&M University

Abstract

Multi-dimensional array is a basic data structure that has been widely used in scientific computing, as well as in many big data analytics applications. Distributed multi-dimensional array has been well studied in the High Performance Computing (HPC) platforms, however, little study has been done in the widely-used big data analytics platforms. In this paper, we present an implementation of Distributed Multi-dimensional Array Toolkit (DMAT) on top of the Apache Spark big data analytics platform. The toolkit supports different ways of multi-dimensional array distributions, repartition, transpose, access, and data parallelism with a variety of parallel execution templates. This paper introduces the software architecture and implementations of DMAT, and also studies the performance characteristics of some typical multi-dimensional array operations with different configurations.

"In situ Analytics of High Frame-rate Image Streaming"

Chiwoo Park
Assistant Professor, Department of Industrial and Manufacturing Engineering, High Performance Material Institute, Florida State University

Abstract

Many scientific studies have been relying on realtime streams of imagery observations of scientific experiments, e.g., imaging dendrite growth in next generation battery systems and growth and control of nanomaterials. The current practice for exploring image streams is as simple as manually scanning over a small part of the image streams to find interesting scientific events, which typically locate very sparsely in the image streams. However, as the frame rate of scientific image streaming increases up to several millions or trillions per second, such manual scanning is no longer feasible, which becomes a major bottleneck against the exploit of imagery observations for scientific research. In this talk, we present a visual analytics framework for automatically identifying important visual events buried in a large volume of high frame rate image streams, which streamlines necessary subtasks including compressive sensing, realtime object identification and tracking, and frequent pattern analytics.

"Building Near-Real-Time Processing Pipelines with the Spark-MPI Platform"

Nikolay Malitsky
Brookhaven National Laboratory

Aashish Chaudhary
Kitware

M. Cowan, K. Kleese Van Dam
Brookhaven National Laboratory

P. O'Leary, S. Jourdain, M. Hanwell
Kitware

Abstract

Advances in detectors and computational technologies provide new opportunities for applied research and the fundamental sciences. Concurrently, dramatic increases in the three V's (Volume, Velocity, and Variety) of experimental data and the scale of computational tasks produced the demand for new real-time processing systems at experimental facilities. Recently, this demand was addressed by the Spark-MPI approach connecting the Spark streaming platform with the MPI high-performance framework. The paper explores this direction within the context of online ptychographic and tomographic reconstruction pipelines.

"Building Blocks for Adaptive Workflows"

Shantenu Jha - KEYNOTE

Associate Professor of Computer Engineering at Rutgers University

Abstract

Next-generation exascale systems will fundamentally expand the reach of biomolecular simulations and the resulting scientific insight, enabling the simulation of larger biological systems (weak scaling), longer timescales (strong scaling), more complex molecular interactions, and robust uncertainty quantification (more accurate sampling). Solving biological problems that require longer timescales, involve more complex interactions and robust uncertainty quantification will require significant algorithmic improvements that incorporate high-level parallelism and leverage the statistical nature of molecular processes. Interestingly, many such simulation algorithms require adaptive workflows. We argue the need for workflow-systems using a building blocks approach to support adaptive workflows on extreme-scale heterogeneous and dynamic resources. We discuss RADICAL-Cybertools as an implementation of the building block concept, and discuss how RADICAL-Cybertools are being used to support adaptive workflows in biomolecular simulations and high-energy physics.

Bio

Shantenu Jha is an Associate Professor of Computer Engineering at Rutgers University. His research interests are at the intersection of high-performance distributed computing and computational science. Shantenu leads the RADICAL-Cybertools project which are a suite of middle-ware building blocks used to support large-scale science and engineering applications. He collaborates extensive-ly with scientists from multiple domains — including but not limited to Molecular Sciences, Earth Sciences and High-Energy Physics. He was appointed a Rutgers Chancellor's Scholar (2015-2020) and was the recipient of the inaugural Chancellor's Excellence in Research (2016) for his cyberinfrastructure contributions to computational science. He is a recipient of the NSF CAREER Award (2013) and several prizes at SC'xy and ISC'xy. More details can be found at http://radical.rutgers.edu/shantenu

"Closed-loop Autonomous Research Systems (ARES)"

Kristofer Reyes
Assistant Professor, Materials Design and Innovation Department, University at Buffalo

P. Nikolaev, A. E. Islam, K. Decker, B. Maruyama,
Air Force Research Laboratory

W. Powell, X. He
Princeton University

Abstract

The acceleration of scientific research is currently being bolstered on two fronts. First is the development of newer, high-throughput experimentation and characterization techniques that allow us to perform a large number of experiments and generate a correspondingly vast amount of data. Second is in the incorporation of data-driven methods to screen, predict and analyze experimental results. In this talk, we describe our current work in coupling these two efforts in an autonomous research system known as ARES. ARES can execute experiments, use experimental results to iteratively learn, and apply gained knowledge to plan future experiments. In this way, ARES closes the experiment-learning-decision loop, allowing it to achieve objectives in an autonomous manner. We will provide a broad overview of the experimental, learning and planning components of ARES, and present results on a pilot program focusing on carbon nanotube growth experiments.

"Is Automated Materials Design and Discovery Possible?"

Michael McKerns
Research Professor, Institute for Advanced Computational Science, Stony Brook University

Abstract

Can we double the scientific output of the neutron and x-ray beamlines? Limitations in our ability to rigorously validate models of beamline experiments force us to make conservative estimates in predicting the beamtime required for an experiment. A critical step in building models of reality and making predictions is solving a statistical optimization problem. Linear and quadratic optimizers and penalties are a mainstay of data science, and have been popular due to their ability to handle large numbers of dimensions quickly. However, the use of linear and/or quadratic tools can seriously limit the amount and quality of information that can be applied in the inverse problem. One could argue that most real-world problems are probabilistic, high-dimensional, and nonlinear with nonlinear constraints -- thus linear and quadratic tools may not actually be a good choice. Too often, we are forced to solve reduced-dimensional problems that may no longer adequately represent reality, but instead fit within the resource and design limitations of the selected optimizer. These limitations become much more pronounced when attempting to predict structure-property relationships in materials, as problems typically require significant computational resources, are nonlinear, and are often governed by rare-events. This talk will introduce some tools within the `mystic' framework for efficiently solving high-dimensional non-convex optimization problems with nonlinear constraints. We will, in the context of materials discovery, also discuss how `mystic', with the OUQ algorithm, can be used for rigorous model validation, certification, and the design of experiments.

"The Convergence of Extreme Computing and Big Data: From Edge Computing to Exascale"

Peter Beckman - KEYNOTE

Co-director of the Northwestern-Argonne Institute for Science and Engineering at Argonne National Laboratory

Abstract

For decades, the basic architecture of extreme-scale systems has been largely static. In one area of our machine room we have compute nodes and in another area, a large shared file system. A slowly evolving, spartan "HPC Software Stack" links the two pieces. This arrangement is out of step both with today's new architectures and a services-based software infrastructure. Parallel file systems are being replaced by object stores, and NVRAM is available everywhere. Many advanced computational science applications are moving past simple bulk-synchronous programming models, and pursuing programming frameworks to support in-situ analysis, live processing of streaming instrument data, and on-demand software stacks. Computing at the edge, where the data is generated, is needed to support massive sensor arrays. Convergence is coming. We need a more agile system software architecture that can simultaneously support both classic HPC computation and new Big Data approaches. From the low-level operating system to the high-level workflow tools, convergence is moving forward. Are we ready?

Bio

Pete Beckman is the co-director of the Northwestern-Argonne Institute for Science and Engineering. From 2008-2010 he was the director of the Argonne Leadership Computing Facility, where he led the Argonne team working with IBM on the design of Mira, a 10 petaflop Blue Gene/Q. Pete joined Argonne in 2002. He served as chief architect for the TeraGrid, where he led the design and deployment team that created the world's most powerful Grid computing system for linking production HPC computing centers for the National Science Foundation. After the TeraGrid became fully operational, Pete started a research team focusing on petascale high-performance system software, wireless sensors, and operating systems. Pete also coordinates the collaborative research activities in extreme-scale computing between the US Department of Energy and Japan's ministry of education, science, and technology. Pete leads the Argo project for extreme-scale operating systems and run-time software. He is the founder and leader of the Waggle project to build intelligent attentive sensors. The Waggle technology and software framework is being used by the Chicago Array of Things project to deploy 500 sensors on the streets of Chicago beginning in 2016. Pete also has experience in industry. After working at Los Alamos National Laboratory on extreme-scale software for several years, he founded a Turbolinux-sponsored research laboratory in 2000 that developed the world's first dynamic provisioning system for cloud computing and HPC clusters. The following year, Pete became vice president of Turbolinux's worldwide engineering efforts, managing development offices in the US, Japan, China, Korea, and Slovenia. Dr Beckman has a Ph.D. in computer science from Indiana University (1993) and a BA in Computer Science, Physics, and Math from Anderson University (1985).

"Comparative Study of Deep Learning Framework in HPC Environments"

Hamid Reza Assadi
PhD Student, Department of Computer Science, Stony Brook University

B. Chapman
Professor, Department of Computer Science, Stony Brook University

Abstract

Deep learning methods have started to become popular in the scientific computing as well. However, since deep learning frameworks are usually developed by communities and companies that mainly employ commodity hardware, the resulting frameworks are not constructed to benefit fully from hardware features available in HPC environments (e.g. high-bandwidth interconnects, etc.), although some work has been done by research teams in this direction. Hence, adapting such frameworks to HPC environments requires performance optimization of these frameworks and, in some cases, changes in their design. Making it easier to install, execute and maintain the frameworks in HPC environments also requires adjustments to the common HPC software stack (i.e. middleware and tools). We have taken a first step toward understanding the behavior of several popular deep learning frameworks in HPC environments. We are particularly interested to find out how scalable and reliable the multi-node support in these frameworks is. We also studied how much support for HPC-specific features each of the frameworks provides. To ensure a fair comparison, we used identical hardware configuration, middleware and tools to perform all the tests. Moreover, all machine learning code used in these tests (i.e. the test models) use the same algorithms and software design principles. This is to ensure that the test results remain comparable even though the exact same code has not been used for testing all frameworks. Finally, the use of standard datasets throughout the tests help us compare our results with previously done research in this area.

"A Scientific Data Provenance Harvester for Distributed Applications"

Eric Stephan
Pacific Northwest National Laboratory

B. Raju, T. Elsethagen
Pacific Northwest National Laboratory

L. Pouchard, C. Gamboa
Brookhaven National Laboratory

Abstract

Data provenance [1,2] provides a way for scientists to observe how experimental data originates, conveys process history, and explains influential factors such as experimental rationale and associated environmental factors from system metrics measured at runtime. The US Department of Energy Office of Science Integrated end-to-end Performance Prediction and Diagnosis for Extreme Scientific Workflows (IPPD) project [3] has developed a provenance harvester that is capable of collecting observations from file based evidence typically produced by distributed applications. To achieve this, file based evidence is extracted and transformed into an intermediate data format inspired in part by W3C CSV on the Web recommendations [7], called the Harvester Provenance Application Interface (HAPI) syntax. This syntax provides a general means to pre-stage provenance into messages that are both human readable and capable of being written to a provenance store, ProvEn [4,5]. This is particularly useful in situations where: 1) the distributed application at runtime already logs simplified provenance evidence ("scruffy provenance"), 2) computer security policies constrain communication protocols used by monitoring application/system services and file evidence is a viable option, or 3) there are barriers to directly incorporating 3rd party monitoring libraries.

"Improving Large Scale Visual Data Analysis Using Intel Supported Software Defined Visualization Solutions"

James Jeffers

Director, Principal Engineer of Software Defined Visualization Engineering for Intel's Enterprise and Government Group

Abstract

As modeling, simulation, and machine learning data output sizes continue almost exponential growth towards Exascale levels, visualizing that data for more rapid discovery and insight is becoming both more important for analyzing results and much more challenging due to data sizes stressing traditional discrete GPU I/O and Memory subsystems. This talk will discuss the benefits and capabilities of a Visualization community initiative supported by Intel for using the extensive parallelism and processing power provided by Intel Xeon® and Intel® Xeon Phi™ many core processors with their large memory capacity and cluster scalability to break many of the barriers to visual fidelity, interactive performance at scale, and seamlessly supporting "In Situ" processing enabling close to real-time analysis turn-around by supporting computational data processing while simultaneously visualizing the results.

"Parallelizing X-ray Photon Correlation Spectroscopy Software Tools Using Python Multiprocessing"

Sameera K. Abeykoon
Research Associate, Computational Science Initiative (CSI), Brookhaven National Laboratory

M. Lin and K. Kleese van Dam
Brookhaven National Laboratory

Abstract

The third generation synchrotron facilities that are designed to deliver highly intense and bright X-rays beams along with the new area detectors capable of achieving high dynamic ratios and fast frame rates have enabled novel Coherent X-ray scattering experiments. X-ray Photon Correlation Spectroscopy (XPCS) is such a technique that measures nano- and meso-scale dynamics in materials. The scikit-beam Python analysis library developed at the National Synchrotron Light Source (NSLS) II at Brookhaven National Laboratory contains a serial version of the XPCS software tools to perform streaming analysis of the structural dynamics of materials which can be time consuming given the anticipated fast data rates and high image resolutions at NSLS-II. It is essential to parallelize these data analysis tools to achieve the best performance on the available workstations with multi-core processors. In this paper, we report the progress we have made in using the Python multiprocessing to parallelize the time-correlation functions in the XPCS software tools. We will compare the results from different multiprocessing approaches, discuss pros and cons associated with each method.

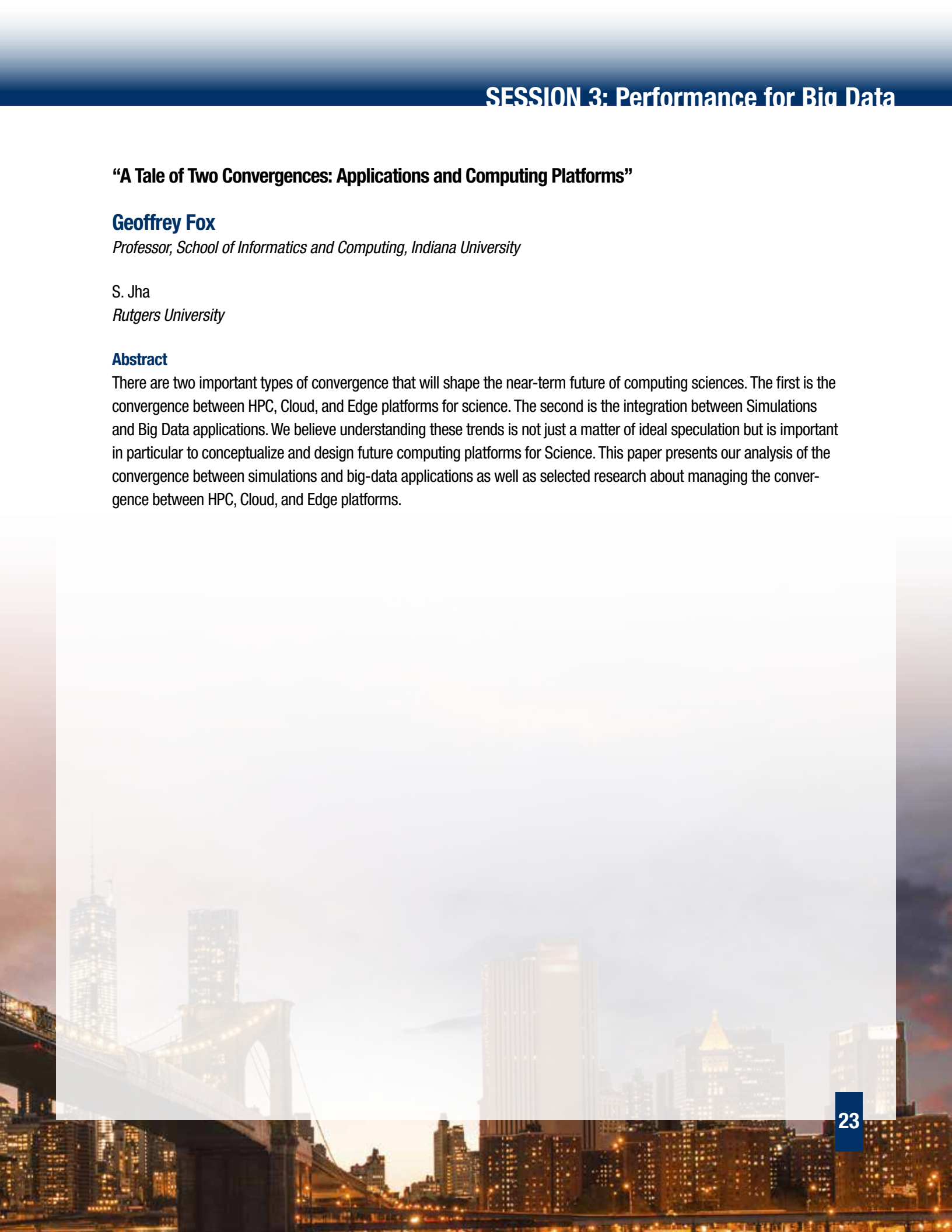
"A Tale of Two Convergences: Applications and Computing Platforms"

Geoffrey Fox
Professor, School of Informatics and Computing, Indiana University

S. Jha
Rutgers University

Abstract

There are two important types of convergence that will shape the near-term future of computing sciences. The first is the convergence between HPC, Cloud, and Edge platforms for science. The second is the integration between Simulations and Big Data applications. We believe understanding these trends is not just a matter of ideal speculation but is important in particular to conceptualize and design future computing platforms for Science. This paper presents our analysis of the convergence between simulations and big-data applications as well as selected research about managing the convergence between HPC, Cloud, and Edge platforms.



"Multigrid Approach for Tomographic Reconstruction"

Zichao (Wendy) Di
Assistant Computational Scientist, Mathematics and Computer Science Division, Argonne National Laboratory

S. Leyffer, S. Wild
Argonne National Laboratory

Abstract

Tomographic imaging refers to the reconstrnction of a 3D object from its 2D projections by sectioning the object, through the use of a penetrating wave from many different directions. This technique requires an accurate image reconstruction; however, the resulting reconstruction problem typically is ill-posed and does not have a unique solution because of insufficient measurements. Different modalities of tomograms have been derived by using different physical phenomena. In particular, X-ray fluorescence (XRF) tomography can be used to reveal the internal elemental composition of a sample while x-ray transmission (XRT) tomography can be used to obtain the spatial distribution of the absorption coefficient inside the sample. To take advantage of the complementary infonnation from different modalities to overcome the ill-posedness, we integrate both modalities which are sinrnltaneously captured during experiment and formulate an optimization approach called joint inversion to simultaneously reconstruct the composition and absorption effect in the sample, Reconstruction performance suffers from the curse of dimensionality. One way of overcoming this challenge is through a multigrid type of method. Multigrid is a general-purpose scalable approach to solving problems with a hierarchy of scales, and it has been successfully applied in many application areas such as vector quantization and image processing. We apply multigrid based optimization framework (MG/OPT) to enhance the reconstruction performance. MG/OPT is designed to accelerate a traditional optimization algorithm applied to a high fidelity problem by exploiting a hierarchy of coarser models. In the context of tomography, we design its hierarchical structure recursively in resolution of the image, as well as the measurement data. We provide several numerical results on the performance of the joint inversion interms of reconstruction quality, as well as significant speedup and improvement of accuracy further provided by MG/OPT. Furthermore , its fast convergence speed with reduced computational cost allows the experimental results to be directly visualized during the experiment at full fidelity and to be used to modify experimental conditions. In particular, the hierarchical nature of our multilevel algorithm allows us to investigate new data acquisition strategies and allow for flexible and adaptive sampling approaches.

"Capturing Provenance as a Diagnostic Tool for Workflow Performance Evaluation and Optimization"

Line Pouchard
Senior Researcher and Applications Architect, Brookhaven National Laboratory

H. Van Dam, W. Xu, A. Malik, K. Kleese Van Dam
Brookhaven National Laboratory

C. Xie
Stony Brook University

Abstract

In extreme-scale computing environments such as the DOE Leadership Computing Facilities scientific workflows are routinely used to coordinate software processes for the execution of complex, computational applications that perform in-silico experiments. Workflows enable the orchestration of the numerous processes that read, write, and analyze data and calculate quantities of interest for parallel and distributed scientific applications that range from quantum chemistry, molecular dynamics, climate modeling, and many others. Monitoring the performance of workflows in HPC provides insights into how a simulation progresses, how the computational resources are used, and where execution bottlenecks occur. But monitoring performance without also simultaneously tracking provenance is not sufficient to understand variations between runs, configurations, versions of a code, and between changes in an implemented stack, and systems, i.e. the variability of performance metrics data in their historical context.

In this talk, we take a provenance-based approach and demonstrate that provenance is useful as a tool for evaluating and optimizing workflow performance in extreme-scale HPC environments. We present Chimbuko, a framework for the analysis and visualization of the provenance of performance articulated around a method for the evaluation of workflow performance that enables the exploration of performance metrics data. The Chimbuko framework captures, analyzes and visualizes performance metrics for complex scientific workflows and relates these metrics to the context of their execution on extreme-scale machines. This is innovative because 1) provenance is used in conjunction with performance measuring tools; 2) this combination is applied to workflows rather than single applications; 3) we provide web-based visualization that enables both high level and detailed visualization of the performance data; 4) provenance metadata is linked to performance metrics; and 5) we demonstrate that provenance is a critical tool for workflow performance evaluation and optimization in extreme-scale environments.

"Event Stream Optimization for Big Data Stream Analytics"

Ryan Quick
Providentia Worldwide

Abstract

Leveraging message fanout and enrichment to provide tailored insight to multiple destinations is an often sought goal for shared data. Solving that problem requires coordinated work in schema design, data durability, and in marshaling. We take an in depth look at several approaches for optimizing events for parallel stream analytics, with a reference implementation and guideline for high performance throughput without sacrificing availability of the workflow or durability of the source and destination data.

"Big Theory for Big Data"

Peter Coveney - KEYNOTE, DINNER SPEAKER
Professor at University College London

Abstract

The current interest in big data and machine learning has generated the widespread impression that such methods are capable of solving most problems without the need for conventional methods of scientific inquiry. One of the characteristic features of big data approaches is that they have a strong "black box" aspect to them, which is at once an advantage and a weakness. The methods are in often presented as being in some sense universal, while at the same time their domain of validity is poorly defined. There is a relative paucity of results available with which to constrain their application in reliable ways while the quantity of data required for the methods to perform reliably is typically underestimated. We look at one or two findings which serve to pinpoint limitations of these approaches as a consequence of the size of the data being investigated. Dynamical phenomena, such as those which commonly arise in life and medical sciences, are particularly challenging for machine learning methods due to the vastness of the data which would need to be acquired in order to apply blind big data methods. Finally, we look at ways in which big data approaches can be made to work synergistically with complementary modelling methods which take into account the structural characteristics of the problem in hand, as exemplified by cancer and the human immune response.

Bio

Prof Peter V. Coveney holds a chair in Physical Chemistry, is an Honorary Professor in Computer Science at University College London (UCL) and is Professor Adjunct at Yale University School of Medicine (USA). He is Director of the Centre for Computational Science (CCS) at UCL. Coveney is active in a broad area of interdisciplinary research including condensed matter physics and chemistry, materials science, as well as life and medical sciences in all of which high performance computing plays a major role. He has published more than 400 scientific papers and co-authored two best-selling books (The Arrow of Time and Frontiers of Complexity, both with Roger Highfield) and is lead author of the first textbook on Computational Biomedicine (Oxford University Press, 2014). Coveney is a founding member of the UK Government's E-Initiative Leadership Council and a Medical Academy Nominated Expert to the UK Prime Minister's Council for Science and Technology on Data, Algorithms and Modelling which has led to the creation of the London based Turing Institute.

"Supercharging the Scientific Process Via Data Science at Scale"

James Ahrens - KEYNOTE

Senior Scientist at Los Alamos National Laboratory

Abstract

Historically, the scientific process is used to explain phenomena by iteratively formulating a theory and running real-world experiments to test and improve the theory. Advances in the field of computer engineering, driven by Moore[1]s law has fundamentally changed the scientific process in two ways:

1. The first change is the availability of inexpensive but highly accurate sensors composed of integrated circuits. The sensors, such as extremely high resolution cameras and signal recorders, enable the collection of scientific data and are used in all scientific disciplines.
2. The second change is the addition of highly detailed scientific simulations that run on high performance computing (HPC) platforms. The performance of these HPC platforms has increased by approximately six orders of magnitude over the past two decades from terascale (10^{12} Floating Points Operations Per Second (FLOPS)) to petascale (10^{15} FLOPS). This performance increase has enabled the creation of extremely detail scientific simulations. These simulations augment the scientific process by providing a proving ground for theories and an environment for virtual experimentation.

Both changes produce massive data streams that need to be effectively processed, transformed, analyzed, visualized and understood through data science. In this talk, I will present new developments in data science, highlighting how the data analysis and visualization process needs to change for exascale supercomputers (10^{18} FLOPS). Exascale supercomputers are bounded by power and storage constraints. These constraints require us to transition from standard, storage-based, post-processing, data science approaches to intelligent, automated, streaming, in situ ones. I will present a new approach that focuses on automatically identifying and tracking areas of interest, and then on selecting and visually presenting these areas to scientists. The work will be presented in the context of solving real-world data science problems for the climate, cosmological, asteroid impact and experimental material science communities.

Bio

Dr. James Ahrens is a senior research scientist at the Los Alamos National Laboratory (LANL). He is the founder and design lead of ParaView, a widely adopted visualization and data analysis package for large-scale scientific simulation data (http://paraview.org). ParaView has had an extremely positive impact on the large-scale data analytic capabilities available to simulation scientists around the world. Dr. Ahrens graduated in 1989 with a B.S. in computer science from the University of Massachusetts and in 1996 with a Ph.D. in computer science from the University of Washington. At LANL, he is part of a data science team of twenty staff, postdocs and students. He is also a national leader of programmatic initiatives important to the United States Department of Energy's National Nuclear Security Administration and Office of Science. Dr. Ahrens is the Data Analysis and Visualization lead for the U.S. Exascale Computing Project and the general chair for this year's IEEE Scientific Visualization conference to be held in Phoenix, AZ in early October.

"Searching for Millions of Objects in the BOSS Spectroscopic Survey Data with H5Boss"

Jialin Liu
HPC Data Analytics Engineer, National Energy Research Scientific Computing/Lawrence Berkeley/Brookhaven National Laboratory

D. Bard
National Energy Research Scientific Computing/Lawrence Livermore/Brookhaven National Laboratory

Q. Koziol, Prabhat
National Energy Research Scientific Computing/Lawrence Berkeley/Brookhaven National Laboratory

Stephen Bailey
Lawrence Berkeley/Brookhaven National Laboratory

Abstract

Spectroscopic surveys of the sky give information about composition of stars and galaxies, and can be used to obtain their redshift, i.e., how fast an object is moving away from the earth. By conducting a redshift survey of many galaxies in the sky, cosmologists can reconstruct the expansion history of the Universe and understand the parameters that describe Dark Energy. Etc. Spectroscopic surveys such as BOSS, the Baryon Oscillation Spectroscopic Survey from the Sloan Digital Sky Survey (SDSS), typically produce a single data file per object observed in the FITS format. This adds up to millions of files, and in the future will be hundreds of millions of files. This is a severe data management problem. One way to alleviate this problem is to covert the many small FITS files to a HDF5 data structure. This will help the metadata problems associated with querying many small files on disk, and will also enable more sophisticated parallel access patterns using the well-developed HDF5 10 libraries. The use cases tested in this work are several typical queries submitted by an astronomer to the dataset, pulling out the information associated with a thousand spectra of some objects of scientific interest, for example quasars. Inthis work, we developed h5Boss, a HDF5 based python package for parallel processing and searching millions of objects in Boss data. The proposed optimal HDF5 structure is beneficial to the broad astronomy community as an alternative file format to FITs. The H5Boss python package offers lightweight query interface and leverages underlying parallel HDFS 1/0 interface for fast data analysis. This study is the first work that designs the file structure based on analytics pattern for BOSS data, and develops the query function with considerations of both metadata overhead and raw data 110 access pattern in BOSS survey analysis. The designed HSBoss is able to scale to 1.6 million fiber objects on 1600 processes on the world's #5 most powerful supercomputers , Cori.

"Robust and Scalable Deep Learning for X-ray Synchrotron Image Analysis"

Dantong Yu
New Jersey Institute of Technology, Brookhaven National Laboratory

Z. Guan, N. Meister
Stony Brook University, Centennial High School

H. Qin
Centennial High School

R. Lashley, B. Sun
Lincoln University

J. Wang, J. Lhermitte, K. Yager
Brookhaven National Laboratory

J. Liu
New Jersey Institute of Technology

Abstract

X-ray scattering is a key technique in modern synchrotron facilities towards material analysis and discovery via structural characterization at the molecular scale and nano-scale. Image classification and tagging play a crucial role in recognizing patterns, inferring meaningful physical properties from sample, and guiding subsequent experiment steps. We designed a deep-learning based image classification software and gained significant success in term of accuracy and speed. Constrained by available computing resources and optimization library, we have to make trade-off among computation efficiency, input image size and volume, and the flexibility and stability of processing images with different levels of qualities and artifacts. Consequently, our deep learning framework requires careful data preprocessing techniques to down-sample imag-es and extract true image signals. However, X-ray scattering images contain different levels of noise, numerous gaps, rotations, and defects arising from detector limitations, sample (mis)alignment, and experimental configuration. Traditional methods of healing x-ray scattering images make strong assumptions about these artifacts and require hand-crafted procedures and experiment meta-data to de-noise, interpolate measured data to eliminate gaps, and rotate and translate images to align the center of samples with the center of images. These manual procedures are error-prone, experience-driven, and isolated from the intended image prediction, and consequently not scalable to the data rate of X-ray images from modern detectors. We aim to design a deep-learning based image classification tool that is fault-tolerant, avoids brute-force down-sample, automates these labor-intensive data preprocessing tasks, and integrates them seamlessly into our TensorFlow based experimental data analysis framework. In particular, we investigate the sensitivity of our TensorFlow based deep learning under different types of arti-facts and propose corresponding mitigating solutions to each type of the artifact:

1) Test the deep learning framework with images that contains different levels of noise, gaps and defects and identify failure scenario.
2) Investigate the correlation between image resolution and prediction accuracy, and make intelligent decision and joint optimization among different aspects of deep learning, i.e., image resolution, prediction accuracy, and the latency of training and inference.
3) Enhance training datasets to comprehensively cover various real-experiment conditions and identified failure scenarios.
4) Leverage distributed Google Tensorflow over multiple GPUs to mitigate the computational cost that is incurred by the enlarged training dataset of high-resolution images.
5) Design rotation-invariant image descriptors to ensure deep-learning to recognition patterns under arbitrary orientation.

1. **Drug Effect Discovery on the Brain Imaging Data**
   Allen Liu, Bryant Liu, *Rocky Point High School*
   J. Cha, Columbia University,
   S. Yoo, Brookhaven National Laboratory

2. **Sensor Network-based Wind Field Estimation Using Deep Learning**
   Daniel Lee, *Commack High School*
   Shinjae Yoo, *Brookhaven National Laboratory*
   D. Cisek, Stony Brook University

3. **Remote sensing data integration for mapping glacial extents**
   Daniel Cisek, *Stony Brook University*
   M. Mahajan, M. Brown, D. Genaway, Stony Brook University

4. **A Transfer Learning approach to parking lot classification in aerial imagery**
   Daniel Cisek, *Stony Brook University*
   Y. Lin, S. Pepper, S.Yoo, Brookhaven National Laboratory
   J. Dale, University of Pennsylvania,
   M. Mahajan, Stony Brook University

5. **Developing and Analyzing OpenMP Benchmarks for GPU Unified Memory**
   Alok Mishra, *Stony Brook University, Brookhaven National Laboratory*
   L. Li, Brookhaven National Laboratory
   B. Chapman, Stony Brook University, Brookhaven National Laboratory

6. **Meeting the Challenges of Data Analysis on the Wire**
   Alya Boumiza, *Baruch College*
   Cole Lewis, Adam Martin, *South Plains College*
   S. Bhattacharyya, J. Zhang, Stony Brook University
   D. Katramatos, M. Yue, S. Yoo, Brookhaven National Laboratory

7. **Network Lasso with L1 and Elastic Net Norms**
   Avinash Barnwal, *Stony Brook University*

8. **Determining the space group of a structure from its atomic pair distribution function**
   Chia-Hao Liu, *Columbia University*
   D. Hsu, Columbia University
   S.J.L. Billinge,Columbia University, Brookhaven National Laboratory

9. **PySHED: a Python framework for Streaming Heterogeneous Event Data**
   Christopher J. Wright, *Columbia University*
   D. B. Allen, J. Lhermitte, M.D. Rocklin, S.J.L. Billinge, Brookhaven National Laboratory

10. **State of Art Numerical Model for Simulating and Forecasting Solar Radiation and Clouds and Data Analysis**
    Mingshen Chen, *Stony Brook University, Brookhaven National Laboratory*
    Y. Liu, S. Yoo, Brookhaven National Laboratory
    X. Li, Stony Brook University

11. **Robust Deep Learning for X-ray Synchrotron Image Analysis**
    Nicole Meister, *Stony Brook University, Centennial High School*
    Z. Guan, H. Qin, Centennial High School
    K. Yager, D. Yu, Brookhaven National Laboratory

12. **Visualization of Higgs potentials and decays from sources beyond the Standard Model including dark matter and extra dimensions**
    Raffaele Miceli, *Stony Brook University*
    M. McGuigan, Brookhaven National Laboratory

13. **Performance Analysis of Hadoop 3.0.0-alpha3 by Running Benchmarks in a Virtual Machine**
    Rohit G. Masur, *New York University*

14. **Spatio-temporal learning from simulation, experimentation, and observation**
Shinjae Yoo, *Brookhaven National Laboratory*
J. Xu, Y. Zhang, X. Zhou, Stony Brook University
M. Chen, Stony Brook University, Brookhaven National Laboratory
D. Lee, Commack High School
B. Feng, Ward Melville High School
G. Lai, W. C. Chang, Y. Yang, Carnegie Mellon School of Computer Science
M. Yue, D. Katramatos, N. D'Imperio, Y. Liu, Brookhaven National Laboratory

15. **Automated Synchrotron X-Ray Diffraction of Irradiated Materials**
John Rodman, *Syracuse University, DOE SULI/BNL*
Y. Lin, D. Sprouster, L. Ecker, S. Yoo, Brookhaven National Laboratory

16. **Efficient full-stack web development for multidimensional scientific data visualization**
Miguel Rodriguez, *SUNY Oswego, Brookhaven National Laboratory*
S. Ha, W. Xu, Brookhaven National Laboratory

17. **Time Series Outlier Detection**
Yaqi Zhang, *Stony Brook University, Brookhaven National Laboratory*
S. Yoo, Y. Liu, Brookhaven National Laboratory

18. **Robust Extreme-scale Multimodal Structured Learning from Spatio-temporal Data --- Earth Science Datasets**
Yangang Liu, *Brookhaven National Laboratory*
S. Yoo, N. D'Imperio, Brookhaven National Laboratory
Y. Yang, Carnegie Mellon University

19. **Near Real Time ETEM Streaming Video Analysis**
Yuewei Lin, *Brookhaven National Laboratory*
S. Yoo, D.N. Zakharov, E. A. Stach, Brookhaven National Laboratory
R. Mégret, University of Puerto Rico

20. **Automated nuclear data validation and verification using critical nuclear reactor models**
Sukhjinder Singh, *Rensselaer Polytechnic Institute*
D. Brown, Brookhaven National Laboratory

21. **Visualization of Higher Genus Carbon Nanomaterials: Free Energy, Persistent Current and Entanglement Entropy**
Tri Duong, *University of Houston, Brookhaven National Laboratory*
M. McGuigan, Brookhaven National Laboratory

22. **Automated data generation for lattice gauge theory**
Daniel Hackett, *University of Colorado, Boulder*
V. Ayyar, W. Jay, University of Colorado, Boulder
E. Neil, University of Colorado, Boulder, Brookhaven National Laboratory

23. **3D Large Eddy Simulation Modeling: Multiscale Clouds and Unsettled Physics**
Xin Zhou, *Stony Brook University, Brookhaven National Laboratory*
Y. Liu, S. Yoo, Brookhaven National Laboratory

24. **Keyword Extraction for Document Clustering using Submodular Optimization**
Xi Zhang, *Stony Brook University, Brookhaven National Laboratory*
S. Yoo, Brookhaven National Laboratory
K. Mueller, Stony Brook University, Brookhaven National Laboratory


PIER 17